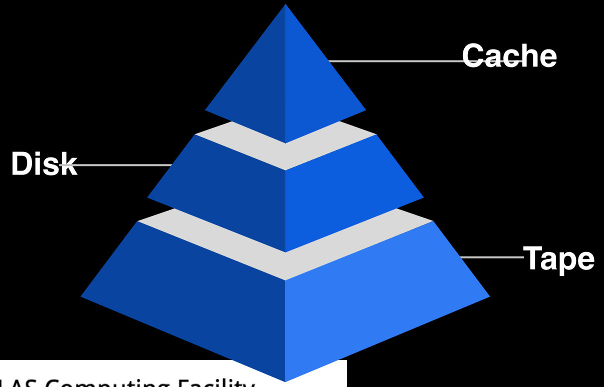
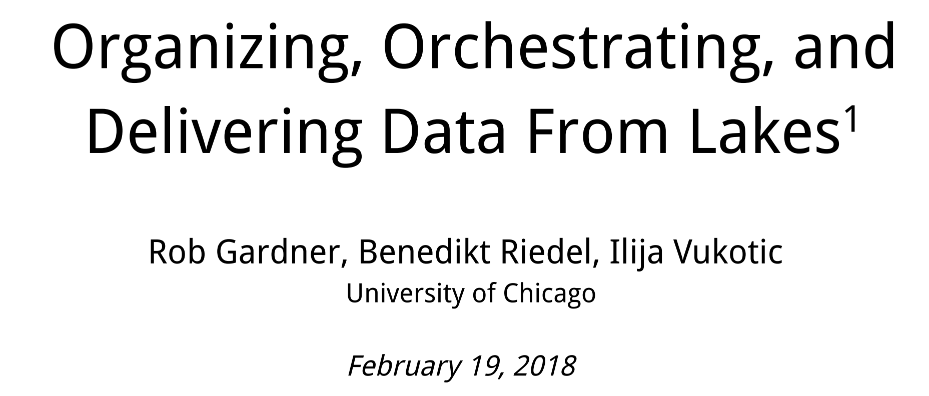
BNL-Lake

A step towards US-Lake

Data lake in the US:
a lot of communication over the last year

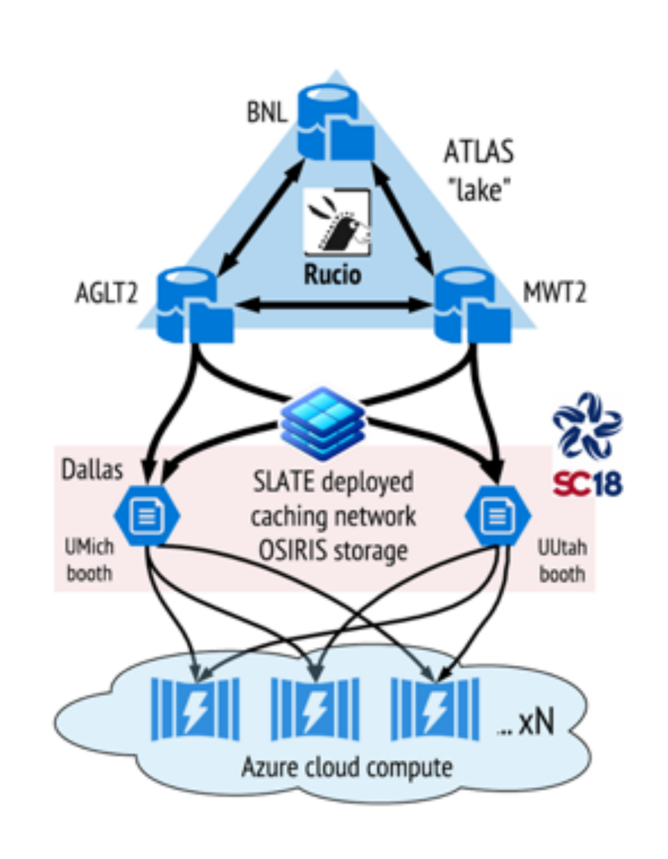


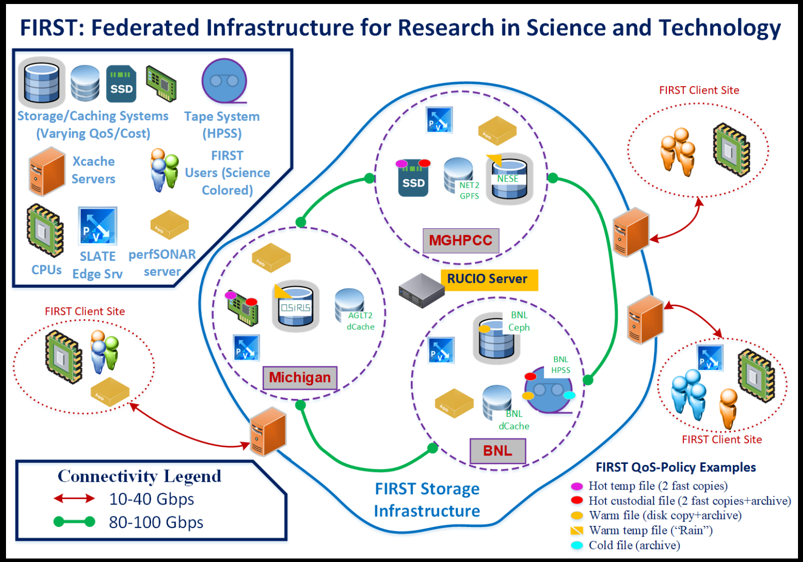


Organizing, Orchestrating, and Delivering Data From Lakes[1]

Rob Gardner, Benedikt Riedel, Ilija Vukotic
University of Chicago

February 19, 2018

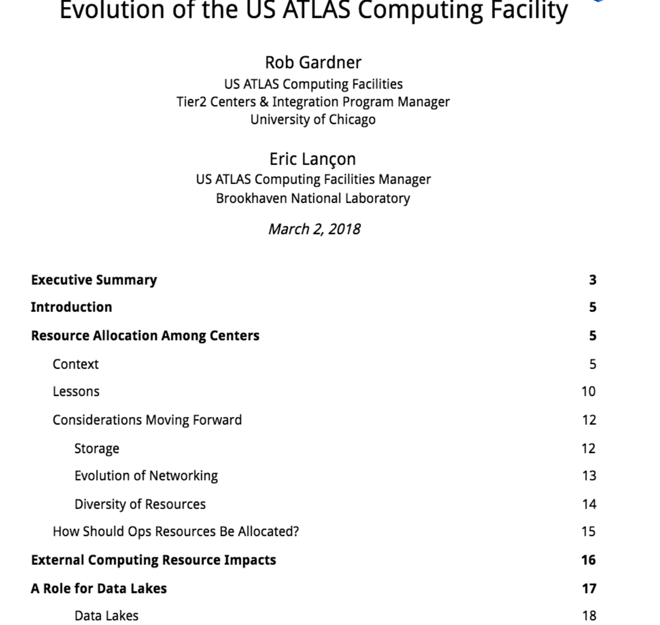




FIRST: Federated Infrastructure for Research in Science and Technology

Connectivity Legend
- ⟷ 10-40 Gbps
- ●—● 80-100 Gbps

FIRST QoS-Policy Examples
- ● Hot temp file (2 fast copies)
- ● Hot custodial file (2 fast copies+archive)
- ● Warm file (disk copy+archive)
- ▨ Warm temp file ("Rain")
- ● Cold file (archive)



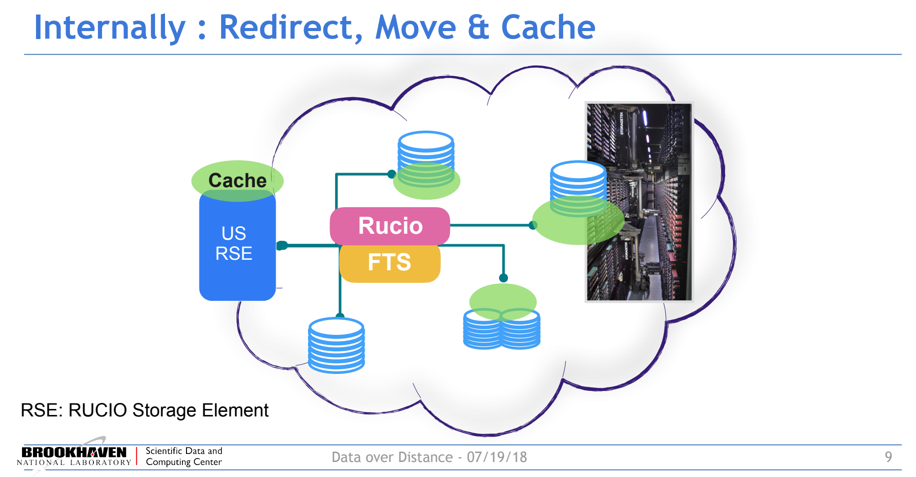
Evolution of the US ATLAS Computing Facility

Rob Gardner
US ATLAS Computing Facilities
Tier2 Centers & Integration Program Manager
University of Chicago

Eric Lançon
US ATLAS Computing Facilities Manager
Brookhaven National Laboratory

March 2, 2018

1



Internally : Redirect, Move & Cache

RSE: RUCIO Storage Element

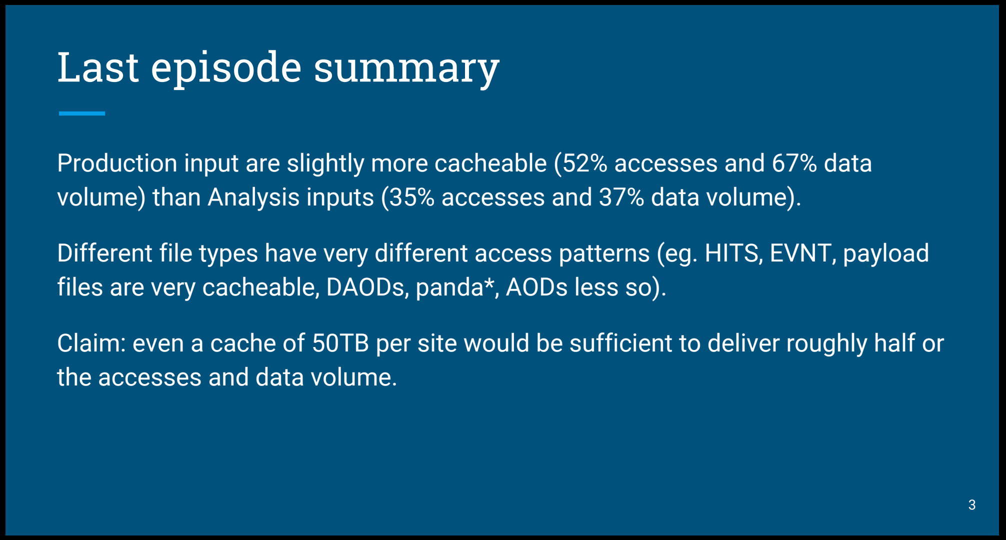
BROOKHAVEN NATIONAL LABORATORY | Scientific Data and Computing Center   Data over Distance - 07/19/18   9

Lake: a 2-D concept

- **Horizontal** dimension: Spatial distribution of storages (the different locations)
- **Vertical** dimension: Hierarchy of storage at a given location
  - Can be very simple (cache only) to complex with a variety of solutions (cache, disk, tape) implementing different technologies
- The 2-D are managed by :
  - An internal file catalog
  - An active information system
- The 2 dimensions can be independently implemented —> **BNL-Lake**

Caching and lake: A note in passing

- Ilija performed extensives studies and simulations (Slides) of caches on US sites
- Ultimate conclusions cannot be reached because of tight connection between data placement (DDM) and workflow management (WFMS)
- Current data organization cannot be used for asserting the benefit of a caching system in a data lake
- Only a real prototype could be used to evaluate the full benefit of a lake



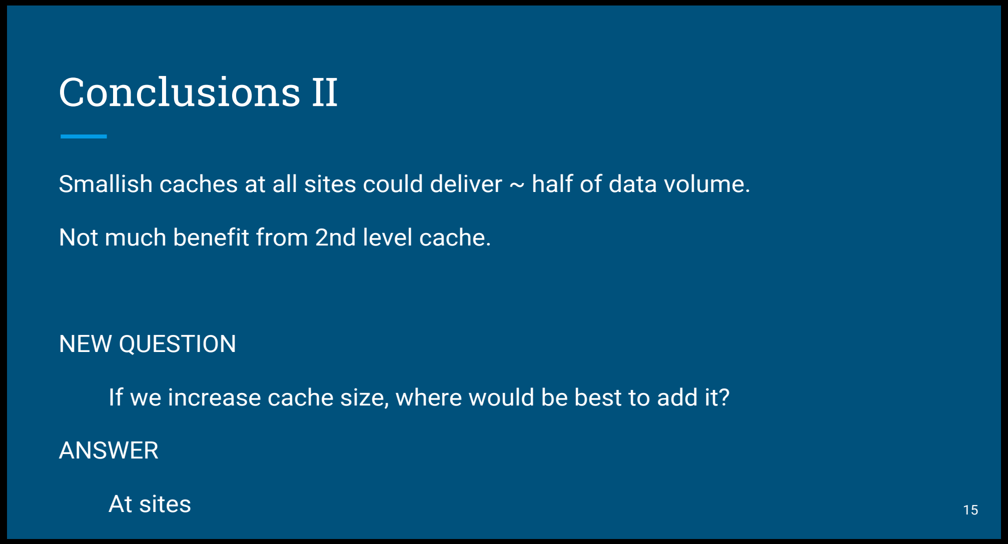
Last episode summary

Production input are slightly more cacheable (52% accesses and 67% data volume) than Analysis inputs (35% accesses and 37% data volume).

Different file types have very different access patterns (eg. HITS, EVNT, payload files are very cacheable, DAODs, panda*, AODs less so).

Claim: even a cache of 50TB per site would be sufficient to deliver roughly half or the accesses and data volume.

3



Conclusions II

Smallish caches at all sites could deliver ~ half of data volume.

Not much benefit from 2nd level cache.

NEW QUESTION

If we increase cache size, where would be best to add it?

ANSWER

At sites

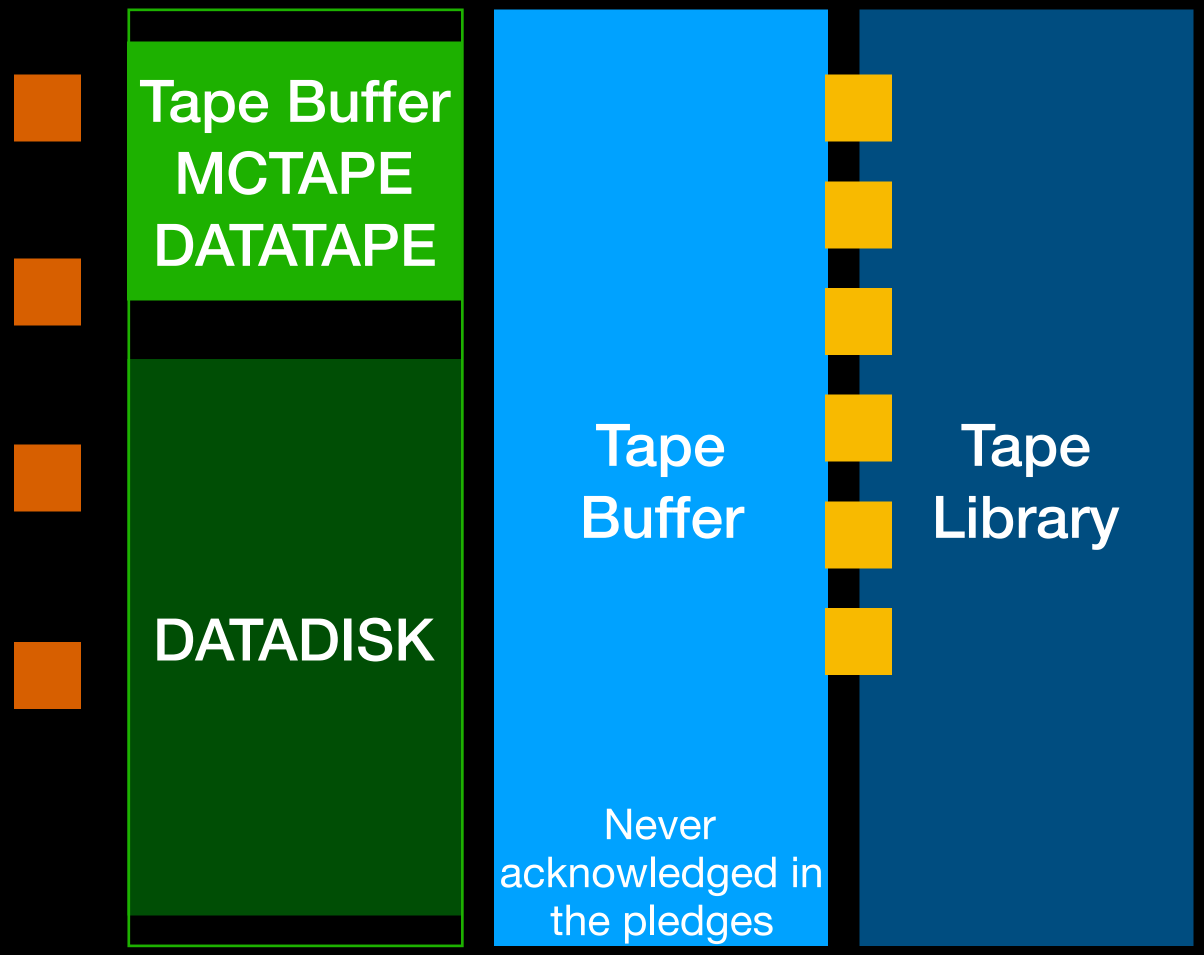
15

Current setup

🟨 Tape Drive

🟧 DTN

🟩 🟩 dCache



Tape Buffer
MCTAPE
DATATAPE

DATADISK

Tape
Buffer

Never
acknowledged in
the pledges

Tape
Library

None of the internals are specified by existing MoUs

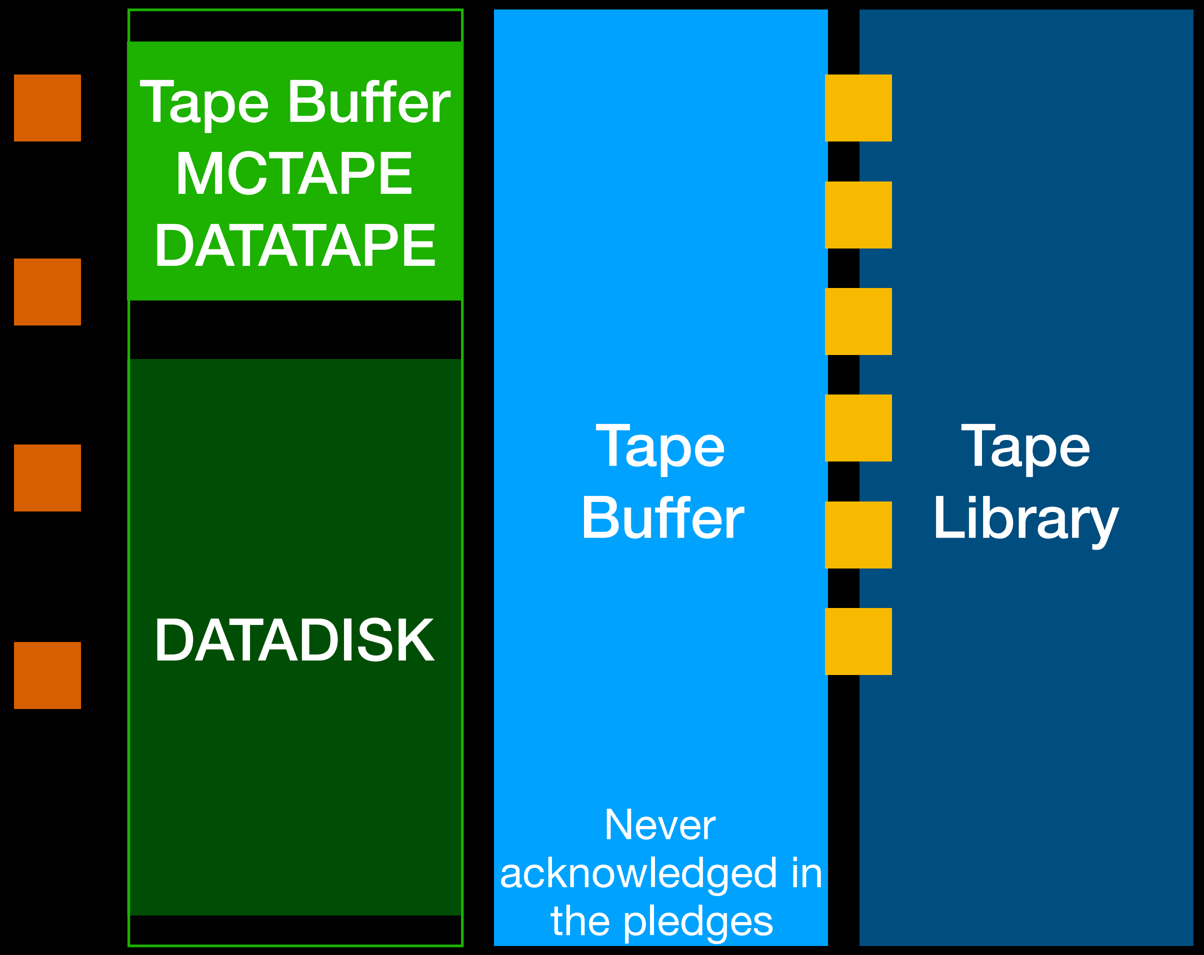
Current setup

🟨 Tape Drive

🟧 DTN

🟩 🟩 dCache

- WFMS/DDM interact with disk tokens and SRM
- The details of site internals are unknown to them (and will remain)
    - Number and type of tape drives
    - Tape buffer size
    - LAN Bandwidth
    - I/O capability and reliability of disk storage,
    - etc… etc… etc…
- Optimisation of data access and storage for given requirements can only be performed by the site



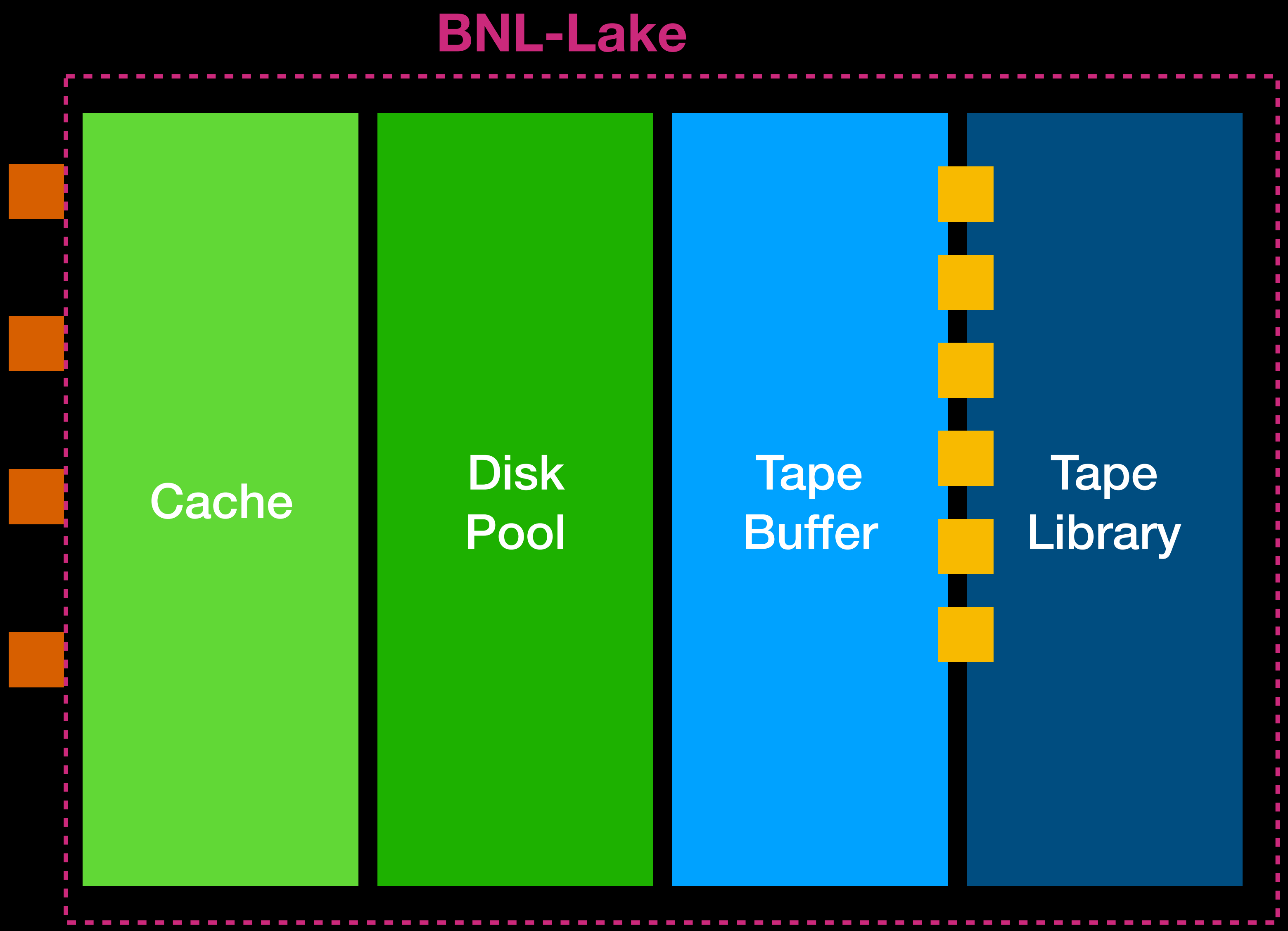
None of the internals are specified by existing MoUs

Proposal

🟨 Tape Drive

🟧 DTN

🟩 🟩 Disk Storage



Proposal

🟨 Tape Drive

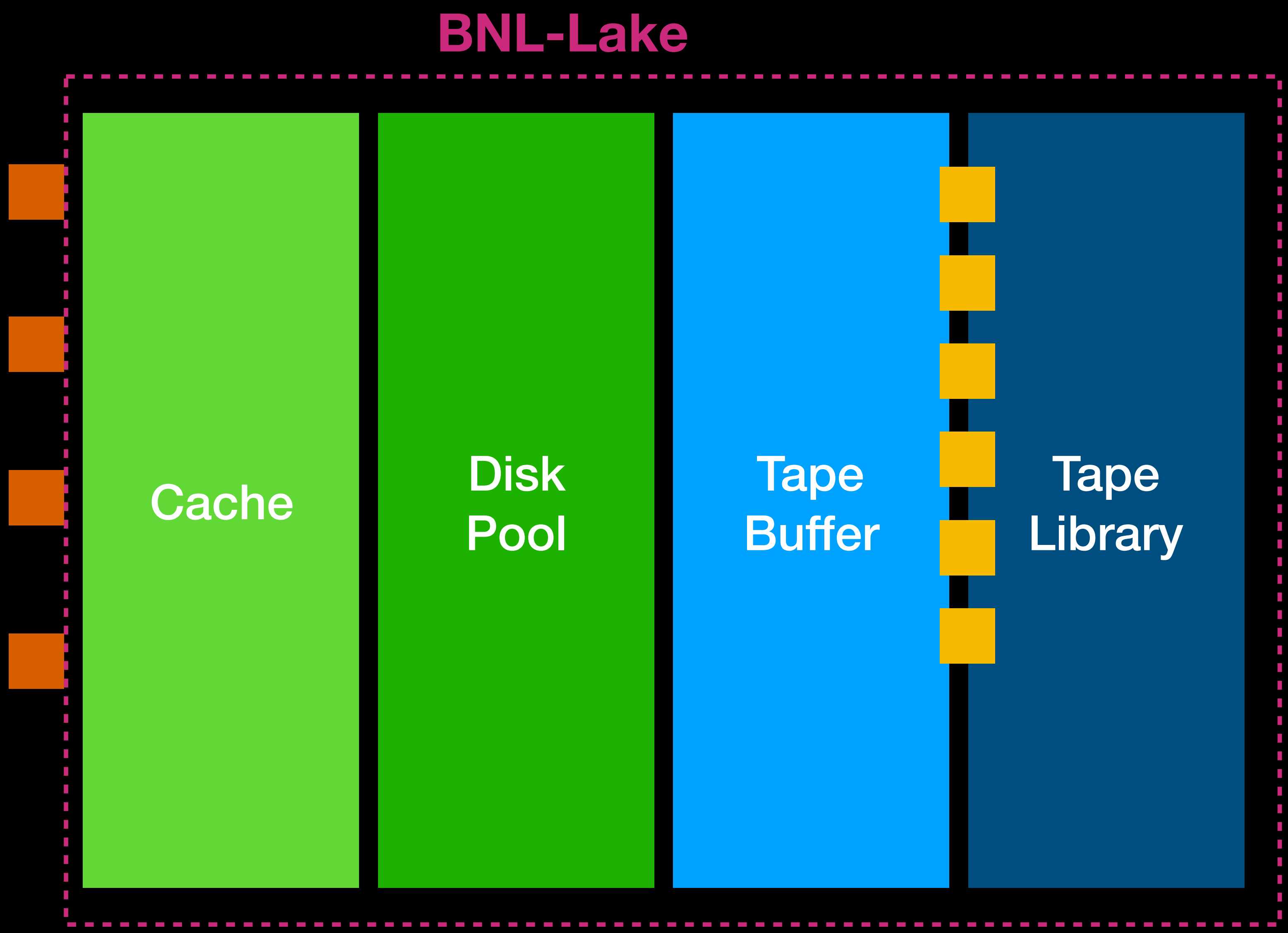
🟧 DTN

🟩 🟩 Disk Storage

- Can be implemented in several steps in parallel to present infrastructure

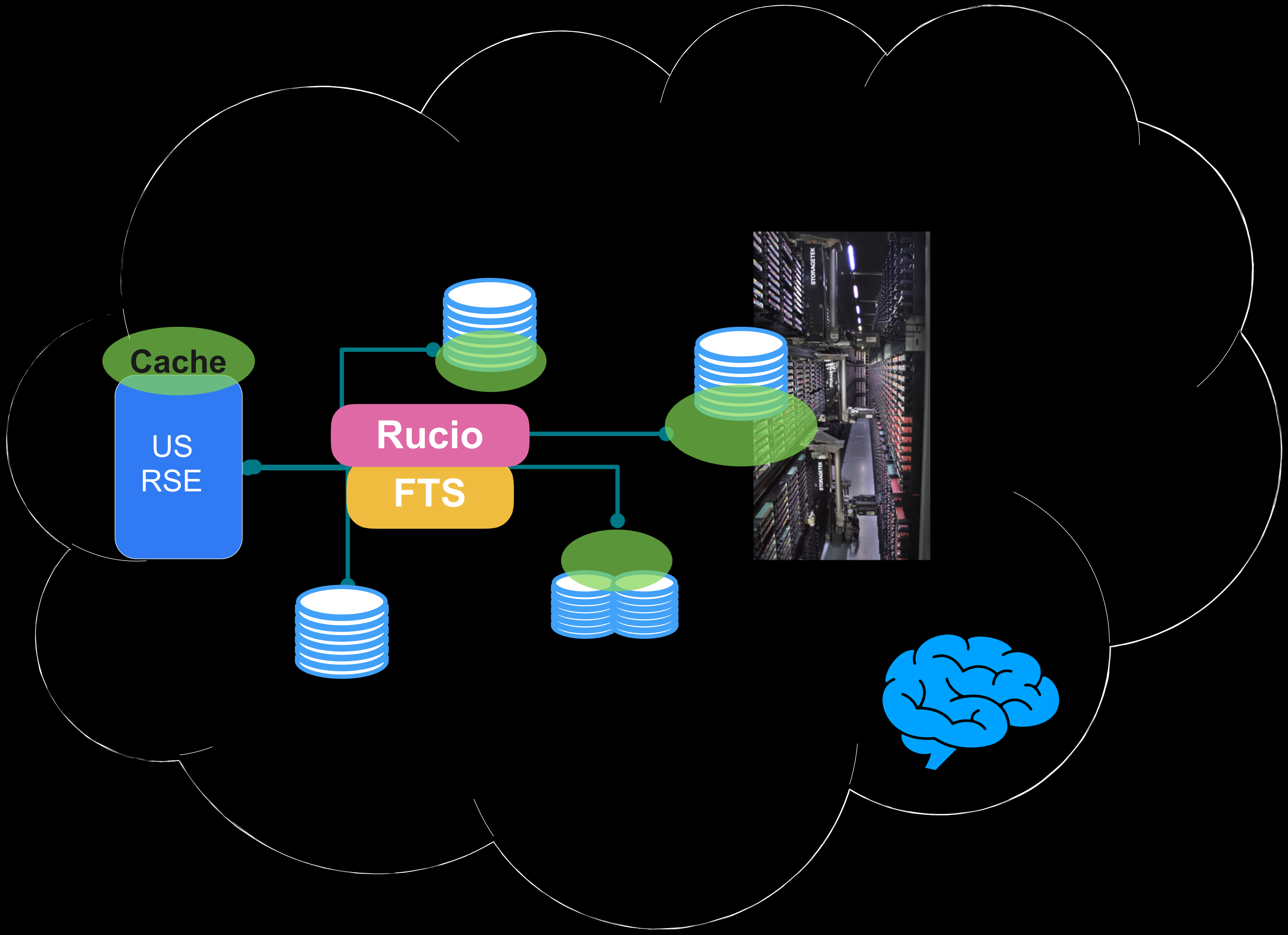
  1. One single disk token
  2. Cache
  3. Information system

- The simplest way to asses the capabilities of the Lake concept without interference of the WFMS/DDM

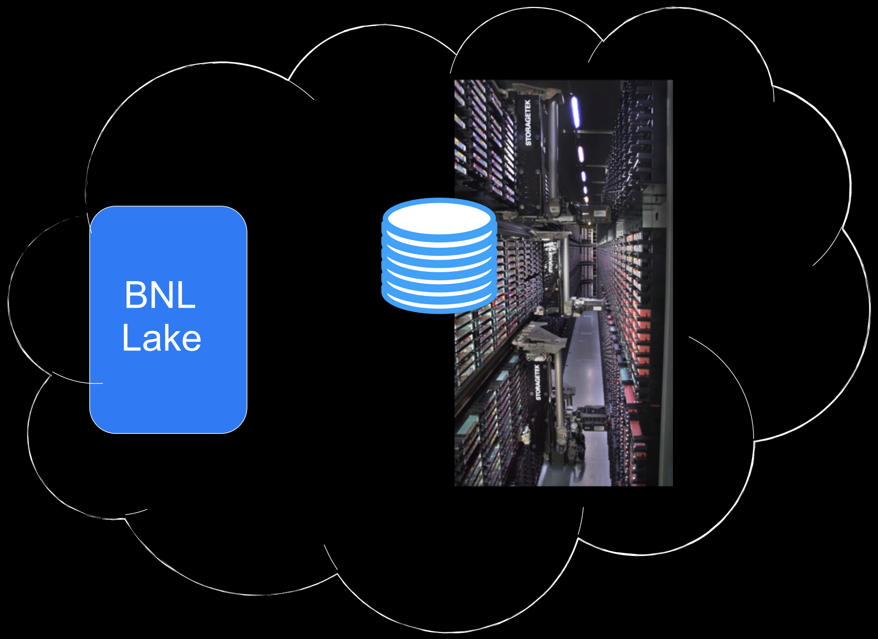


Note: this concept will break the historical WLCG requirement of a given disk space at a given site

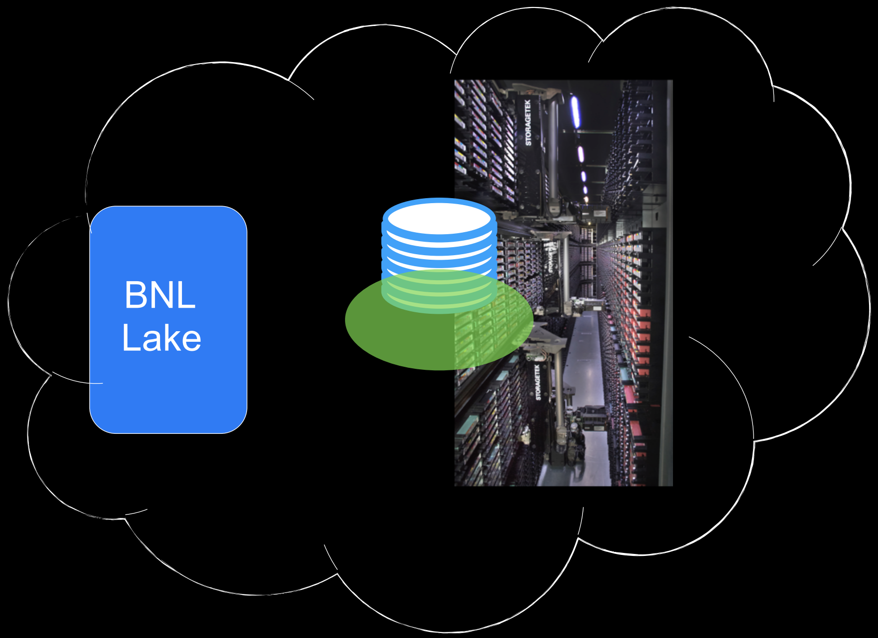
The dream…US-Lake



PathForward

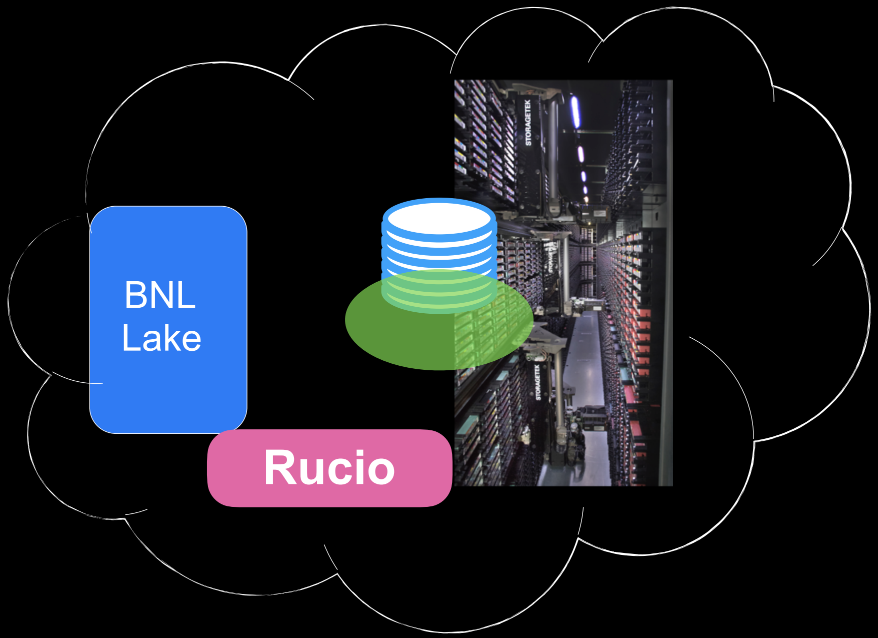


PathForward

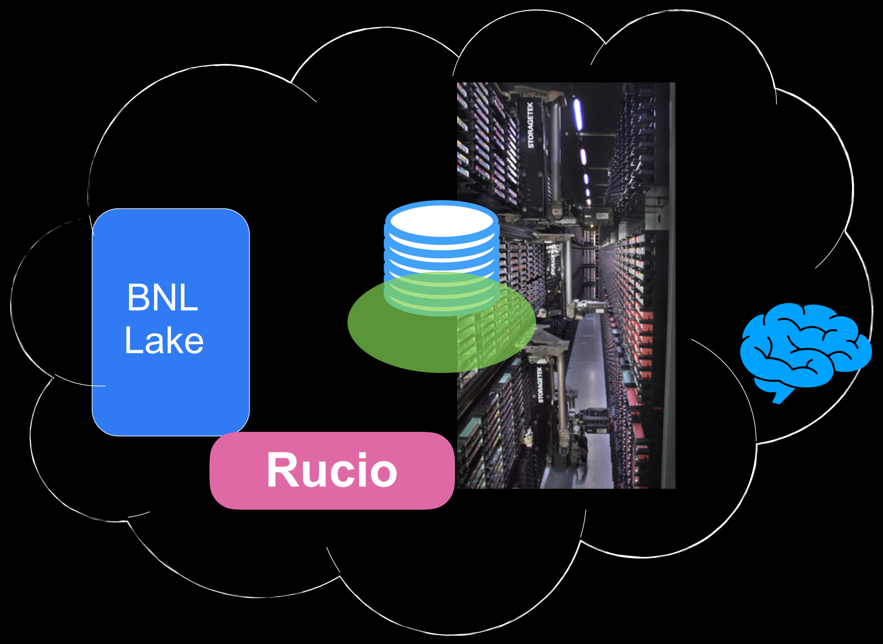


BNL Lake

PathForward



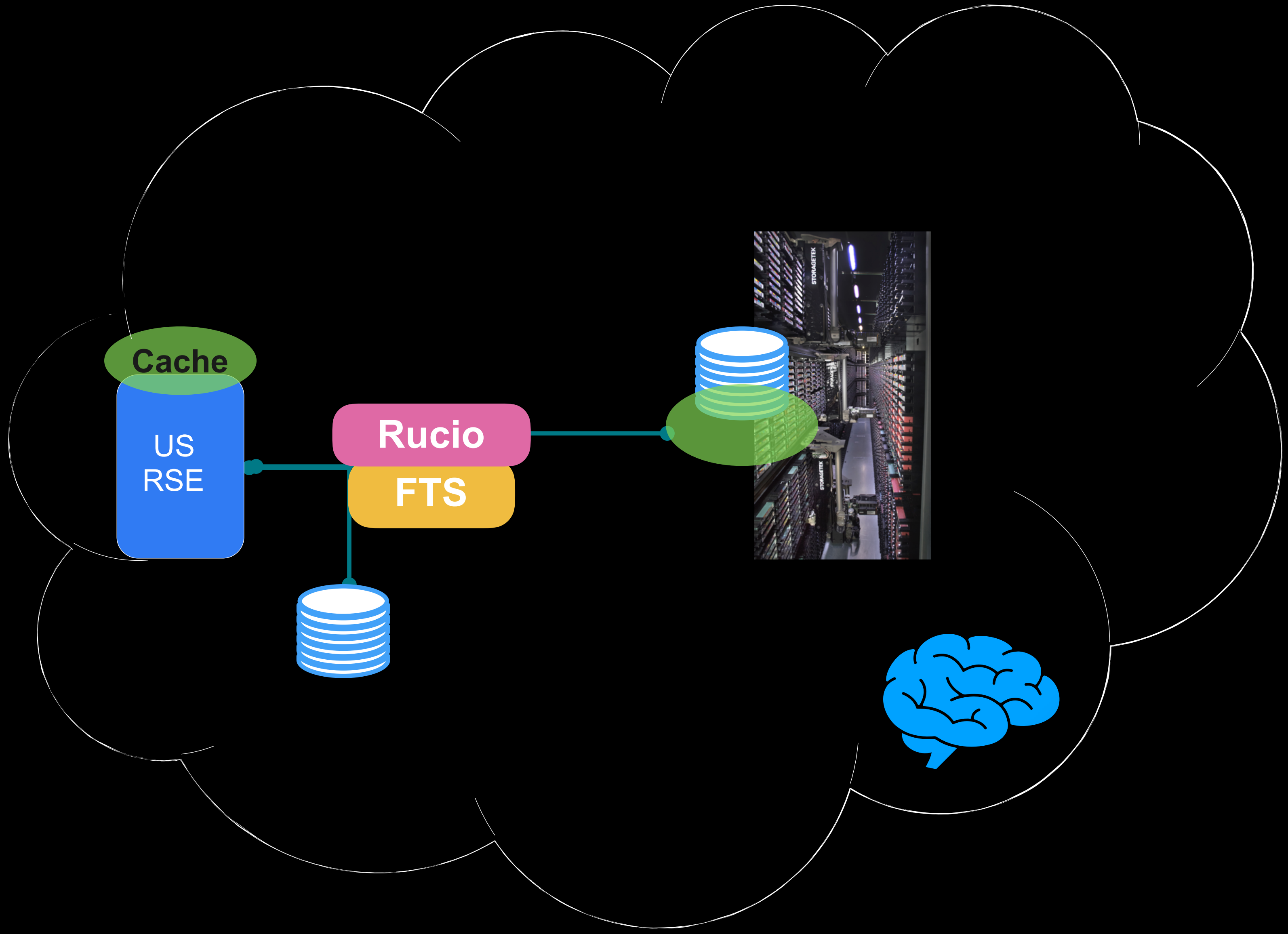
PathForward



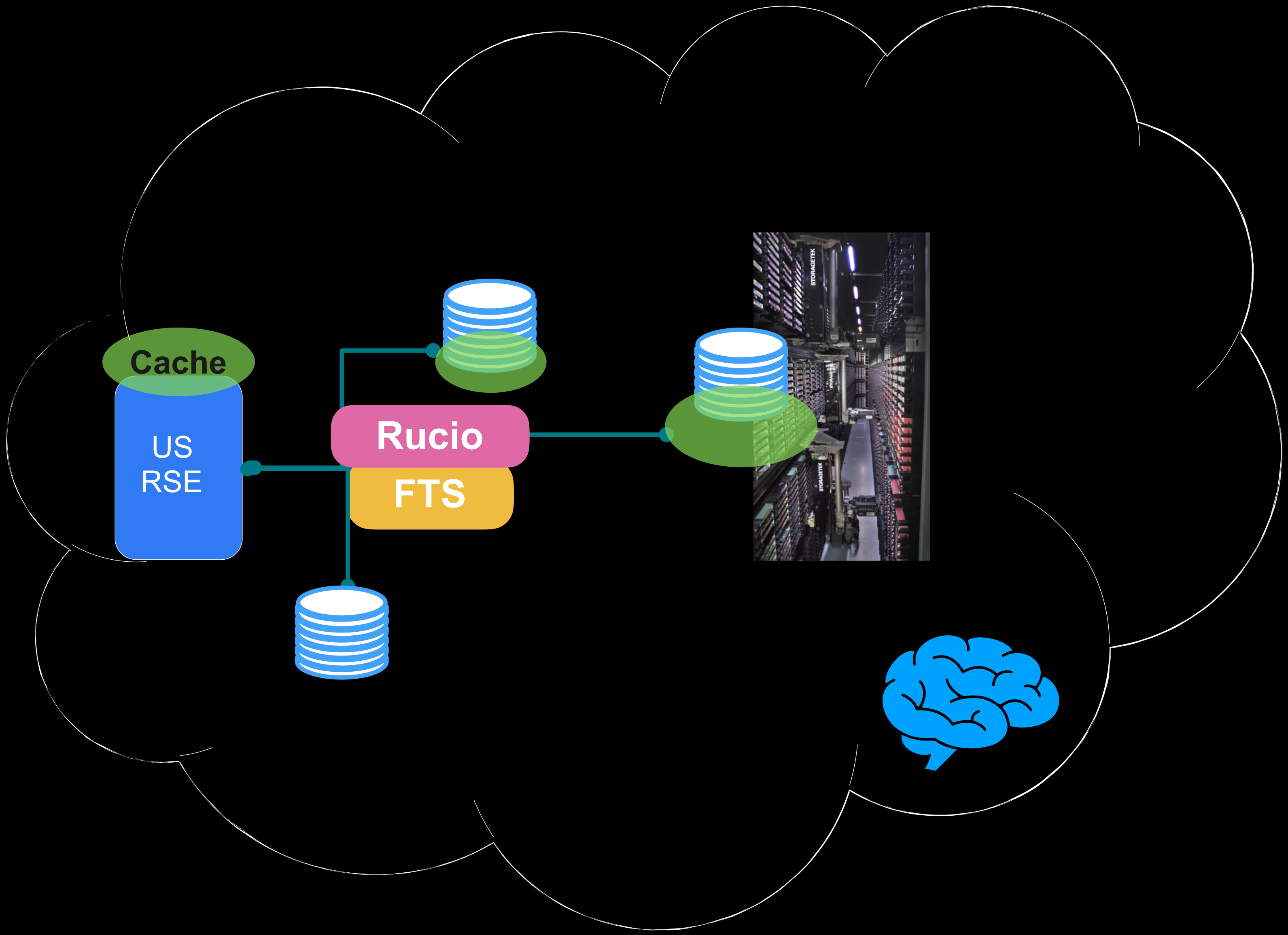
BNL Lake

Rucio

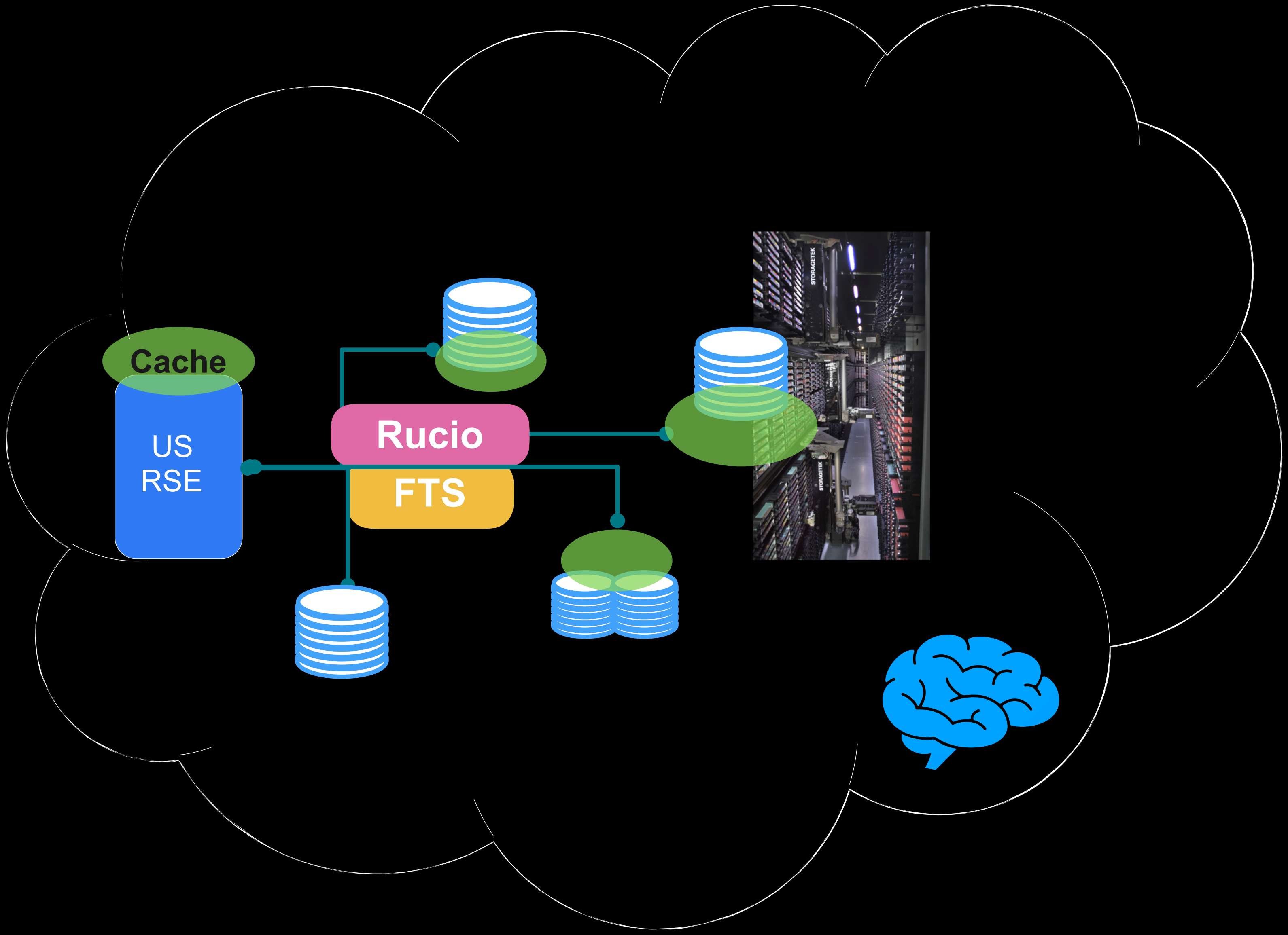
PathForward



PathForward



PathForward


Cache
US
RSE
Rucio
FTS